# Visual analysis of interactive document clustering streams

Eric M. Cabral
cabral.eric@usp.br
Universidade de São Paulo
São Carlos, São Paulo, Brazil

Evangelos E. Milios
eem@cs.dal.ca
Dalhousie University
Halifax, Canada

Rosane Minghim
r.minghim@cs.ucc.ie
University College Cork
Cork, Ireland

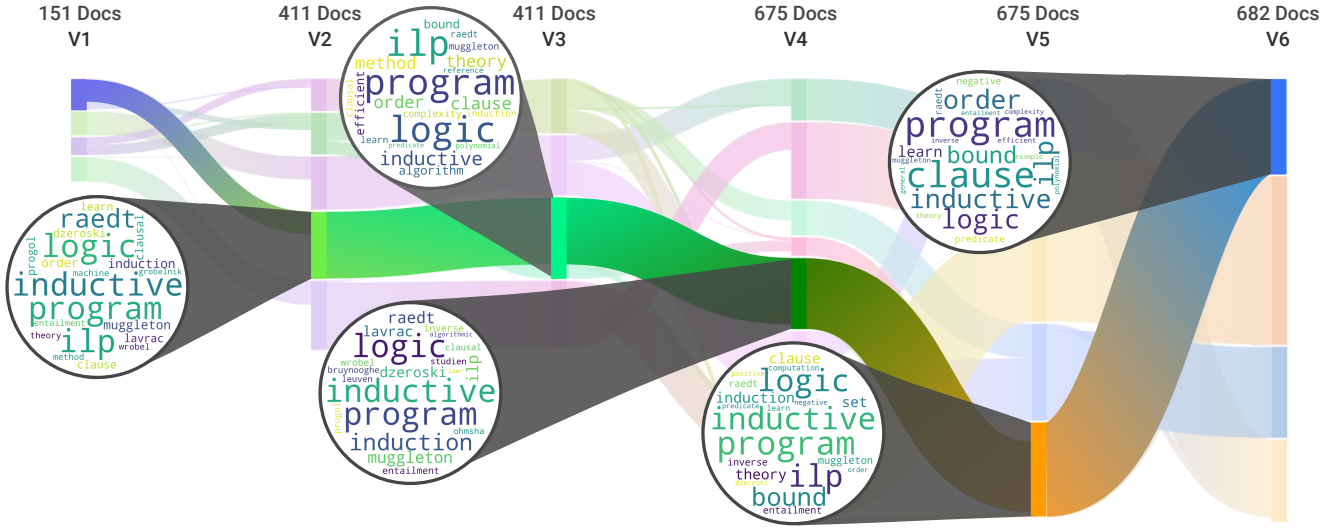


**Figure 1: Layout based on a collection of 682 papers after six clustering iterations. Events in the sequence of steps are both user interactions and increment in data set size. The clusters are highlighted from their rectangular representation. The keyterm cloud in the figure shows a cluster that remains stable even under changes in the corpus.**

## ABSTRACT

Interactive clustering techniques play a key role by putting the user in the clustering loop, allowing her to interact with document group abstractions instead of full-length documents. It allows users to focus on corpus exploration as an incremental task. To explore Information Discovery's incremental aspect, this article proposes a visual component to depict clustering membership changes throughout a clustering iteration loop in both static and dynamic data sets. The visual component is evaluated with an expert user and with an experiment with data streams.

## CCS CONCEPTS

• **Information systems** → **Clustering**; • **Human-centered computing** → **Visual analytics**.


*AVI '20, September 28-October 2, 2020, Salerno, Italy*

ACM ISBN 978-1-4503-7535-1/20/09.
https://doi.org/10.1145/3399715.3399962

## KEYWORDS

Document clustering, Visual analytics, Document streams, Text visualization

**ACM Reference Format:**
Eric M. Cabral, Evangelos E. Milios, and Rosane Minghim. 2020. Visual analysis of interactive document clustering streams. In *International Conference on Advanced Visual Interfaces (AVI '20), September 28-October 2, 2020, Salerno, Italy.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3399715.3399962

## 1 INTRODUCTION

The visual analytics system Vis-Kt [1, 3] proposes an effective keyterm-based document clustering technique where the user interacts with the top keyterms discovered by the clustering algorithm and can tailor the results of the cluster as her understanding of the dataset evolves. By employing keyterms as the basis for content interaction and clustering, the approach becomes intuitive for a varied set of end-users.

Since Information Discovery is an incremental task, Vis-Kt can take advantage of the visualization of interaction history, depicting each clustering structure sequentially. With this kind of visual component, we can depict the unsupervised clustering of temporal data and the the user-guided clustering through interactions. Such functionality was not yet developed for Vis-Kt.

This article adapts a popular work flow visualization technique called Sankey diagram [2] to depict the evolution of clustering

structures within Vis-Kt to allow the user to track her interactions throughout the clustering loop. To evaluate this approach, we first engaged a textual data visualization researcher in an expert study. Second, we experimented with a tailored streaming dataset to evaluate the clustering tracking throughout subsequent steps in the clustering loop. The expert study and the streaming data experiment highlights how the Sankey diagram can help the user to obtain deeper knowledge into the explored corpus by analyzing how the clustering structure evolves.

## 2 THE VISUAL COMPONENT

The Sankey diagram was designed as a Directed Acyclic Graph (DAG). Within the DAG structure, $M$ layers represent the $M$ clustering steps along the clustering loop, which we call "sessions". A cluster associated with a session is represented as a node in the graph. The links between two nodes are defined by the document cluster membership changing from one session to the next.

With the Sankey diagram, the user can track how her feedback affected the clustering result. She can highlight any cluster at any time, explore the documents that are included in the selected cluster and analyze its cluster membership in previous sessions. The top five keyterms of each cluster are shown by hovering over the cluster representation.

The cluster split and merge operations are intuitively represented as flows on the Sankey diagram. In the split operation, a cluster represented by a rectangle has two or more links to other clusters in the next session. In the merge operation, a cluster has two or more links from the previous session. The visual component and the operations are depicted in Figure 2.

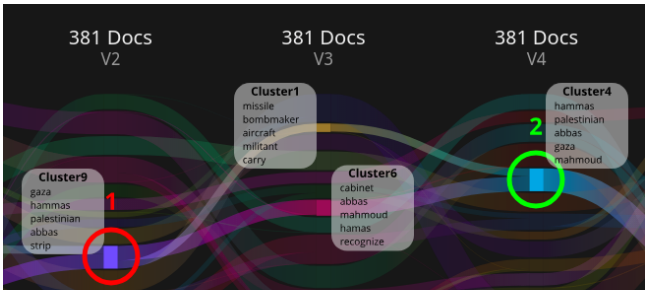


**Figure 2: (1) A split operation of Cluster9 to Cluster1 and Cluster6, (2) a merge operation of Cluster1 and Cluster6 into Cluster4. The top 5 key-terms listed within the gray boxes.**

## 3 USAGE SCENARIOS

In the following sections, we report two usage scenarios that were explored through this research. The first is related to how our adapted Sankey diagram can be used to depict clustering flows throughout several clustering iterations and how the visual component addresses the scenario of data streams. Next, we report an expert study executed to evaluate the usefulness of the proposed visual component for clustering evolution analysis.

### 3.1 Incremental Clustering Analysis

The proposed visual component has the dual purpose of depicting clustering iterations with user feedback to the clustering algorithm and depict text data stream clusters. In both scenarios, the visualization is useful for the user to recognize evolving clustering patterns. If a cluster is nearly unchanged between several executions or iterations, this means that this is a consistent and well-defined cluster.

This clustering behavior is depicted in Figure 1. In this usage scenario, a data collection of 682 documents related to the field of Computer Science is partitioned in six proportionally equal partitions. The clustering at V1, V2, V4, and V6, are examples of incremental clustering, whereas the clustering at V3 and V5 are generated with user feedback. We can notice that both scenarios are seamlessly represented in the Sankey diagram, and the clustering structure shows term consistency throughout five iterations with different types of evolution (due to additional data arriving and due to user interaction with the clustering). This is confirmed with each cluster's WordClouds [4].

### 3.2 Expert Study: Clustering Evolution with User Interactions

To evaluate the proposed visual component, we engaged one visual analytics expert to provide quantitative feedback on the utility of the Sankey Diagram to depict the interactive clustering steps. The expert is a Masters' student with experience in visual analytics on textual data. The expert used a news article dataset of his choice.

The expert found the clustering visualization of the Sankey Diagram more intuitive than the whole set of visual components included in Vis-Kt. He reported that after six clustering interactions, the resulting flow visualization on the Sankey Diagram brought him more insight into the content of the corpus than the previous interaction with Vis-Kt's main visual components. More than just depicting the results of the clustering loop, the expert suggested that the diagram should have a central role in the system, providing clustering interaction tools. For instance, the user may visually choose which clusters she wants to merge or split by dragging and dropping the cluster representations in the Sankey Diagram

## 4 CONCLUSIONS AND FUTURE WORK

The visual approach and experiments conducted in this work contribute towards a better understanding of how evolutionary visualization can enhance the information discovery process by during interactive document clustering. The feedback from this work bring novel and promising perspectives towards embedding the proposed Sankey diagram as a major visual component to support intuitive keyterm-based clustering. Because of its flexible design, it can be scaled to growing datasets and applied to incremental clustering scenarios. The next step shall evaluate the proposed visual component in other domains to improve its validation.

## ACKNOWLEDGMENTS

The authors thank the participant in the expert study. The research was funded by CNPq (Brazil, grants numbers 07411/2016-8 and 133685/2018-7),the Natural Sciences and Engineering Research Council of Canada, The Boeing Company, CALDO, and the International Development Research Center, Ottawa, Canada.

# REFERENCES


[1] S. Nourashrafeddin, E. Sherkat, R. Minghim, and E. Milios. 2018. A Visual Approach for Interactive Keyterm-Based Clustering. *ACM Trans. Interactive Intelligent Systems* 8, 1 (2 2018), 1–35. https://doi.org/10.1145/3181669

[2] P. Riehmann, M. Hanfler, and B. Froehlich. 2005. Interactive Sankey diagrams. In *IEEE Symp. Information Visualization (INFOVIS)*. IEEE, Minneapolis, MN, USA, 233–240. https://doi.org/10.1109/INFVIS.2005.1532152

[3] E. Sherkat, S. Nourashrafeddin, E. Milios, and R. Minghim. 2018. Interactive Document Clustering Revisited: A Visual Analytics Approach. In *Int. Conf. on Intelligent User Interfaces (IUI)*. ACM Press, New York, New York, USA, 281–292. https://doi.org/10.1145/3172944.3172964

[4] F. Viégas and M. Wattenberg. 2008. TIMELINESTag clouds and the case for vernacular visualization. *interactions* 15, 4 (7 2008), 49. https://doi.org/10.1145/1374489.1374501